

Running Head: LONGITUDINAL EFFECTS ON ADOLESCENT LANGUAGE

Longitudinal Effects on Early Adolescent Language: A Twin Study

Nicole Harlaar, Department of Pediatrics, School of Medicine, University of Colorado

Laura Segebart DeThorne, University of Illinois

Jamie Mahurin Smith, Illinois State University

Mariana Aparicio Betancourt, University of Illinois

Stephen A. Petrill, The Ohio State University

Note: Both Nicole Harlaar & Laura DeThorne contributed equally to the present manuscript.

This accepted version of the manuscript does not include minor corrections that were made to the page proofs. The final publication is available at

<http://jslhr.pubs.asha.org/journal.aspx>

Harlaar, N., DeThorne, L.S., Smith, J.M., Betancourt, M.A., & Petrill, S.A. (2016). Longitudinal effects on early adolescent language: A twin study. *Journal of Speech, Language, and Hearing Research*, 59(5), 1059-1073.

Abstract

Purpose: We evaluated genetic and environmental contributions to individual differences in language skills during early adolescence, measured both by language sampling and standardized tests, and examined the extent to which these genetic and environmental effects are stable across time.

Method: We employed structural equation modeling on latent factors to estimate additive genetic (A), shared environmental (C), and nonshared environmental (E) effects on variance in standardized language skills (i.e., Formal Language) and productive language sample measures (i.e., Productive language) in a sample of 527 twins across three time points (mean ages 10-12).

Results: Individual differences in the Formal Language factor were influenced primarily by genetic factors at each age, whereas individual differences in the Productive language factor were primarily due to nonshared environmental influences. For the Formal language factor, the stability of genetic effects was high across all three time points. For the Productive language factor, nonshared environmental effects showed low but statistically significant stability across adjacent time points.

Conclusions: The etiology of language outcomes may differ substantially based on assessment context. In addition, the potential mechanisms for nonshared environmental influences on language development warrant further investigation.

Literature Review

In the child language literature, adolescent development often gets overshadowed by the exciting linguistic milestones of the earlier years. Nevertheless, it is difficult to imagine a more significant transitional period than adolescence. Defined as the maturational period spanning the onset of puberty to early adulthood¹, adolescence is associated with significant neurobiological, physical, cognitive, social, and behavioral changes (APA, 2002; Dahl, 2004; Rice & Barone Jr., 2000; Steinberg, 2005). Given evidence of increased brain plasticity during this period, some have referred to adolescence as a second critical period of development (Dahl, 2004; Huttenlocher, 1994; Peper et al., 2009; Rice & Barone Jr., 2000). In regard to communication-related skills, the adolescent years are marked by growth and increasing individual differences in metalinguistic and abstract thinking, such as use and understanding of figurative language and morphologically complex nouns and adjectives (Nippold, 2000; Nippold & Sun, 2008; Paul, 2001), development of an adolescent register (Nippold, 2000), and increased sophistication in theory of mind, particularly as it relates to affect (Blakemore, 2012). In addition, the period of adolescence has been associated with a narrowing of linguistic flexibility, at least in terms of acquiring native competencies in domains such as phonology and morphosyntax (e.g., Johnson & Newport, 1989; Thompson et al., 2000). The present study examines the stability of etiological effects on language development explicitly during early adolescence.

Longitudinal Twin Studies of Language

Longitudinal twin studies that compare monozygotic (MZ) and dizygotic (DZ) twins are a useful methodology for examining genetic and environmental influences associated with

¹ Given that adolescence is often defined relative to the onset of puberty, the specific age range varies based on race and biological sex, amongst other factors (Herman-Giddens et al., 1997; Marshall & Tanner, 1969, 1970). For the purpose of this paper, we are associating adolescence with an age range of 10-18 years in accordance with the American Psychological Association (2002)

adolescent development. Twin studies offer a natural control for genetic effects because MZ twins are twice as similar genetically as DZ twins. Twin studies also facilitate the estimation of environmental effects, which can be divided into two types: shared environmental factors, reflecting variation in non-genetic influences that contribute to familial resemblance (e.g., family SES); and nonshared environmental factors, reflecting variation in environmental influences that cause individual family members to differ from one another (e.g., different educational experiences). Twin similarity is increased by genetic and shared environmental factors, and reduced by nonshared environmental factors. The genetic effect size is typically referred to as heritability (a^2 or h^2), with shared and nonshared environmental effects represented as c^2 and e^2 , respectively. Estimates of genetic effects on individual differences in language development at different ages have varied widely across studies, with such variability attributed to differences in language domain (e.g., Stromswold, 2001), form of measurement (e.g., DeThorne et al., 2008), environmental circumstances (e.g., Rowe, Jacobson, & Van den Oord, 1999), and child age (e.g., Spinath, Price, Dale, & Plomin, 2004). Specific to child age, multivariate analyses with longitudinal data allow us to examine the stability of genetic and environmental effects over time.

One example of a longitudinal twin analysis, Haworth et al. (2010), combined data across six twin studies from four countries to study potential developmental change in the etiology of individual differences in IQ using a total sample of 11,000 twin pairs. Twins were grouped according to childhood (4-10 years), adolescence (11-13 years), and young adulthood (14-34 years). Across these three developmental periods, estimates of heritability increased linearly from 41% in childhood to 66% in young adulthood, with a middle ground of 55% heritability during adolescence (see also Hoekstra, Bartels, & Boomsma, 2007; Oliver & Plomin, 2007). The

authors note that increased heritability of general cognitive ability could be due to the emergence of new genetic effects or the process of gene-environment correlation (p. 1118). As an example, active gene-environment correlation suggests that as children gain increased control over their environments, they are likely to select experiences and contexts that are more consistent with their own genetic inclinations, thereby predicting increased heritability over time (cf. Hopper, 2000; Plomin, DeFries, & Loehlin, 1977; Scarr & McCartney, 1983).

Specific to language, Hayiou-Thomas, Dale, and Plomin (2012) offered “the first long-term longitudinal examination of the etiology of individual differences in language from early childhood through to adolescence (p.233).” The analyses included approximately 8,000 same-sex twin pairs from the Twins Early Development Study (TEDS) and focused on three time periods: early childhood (2-4 years), middle childhood (7-10 years), and early adolescence (12 years). The study noted an increase in heritability as children moved from early childhood ($a^2=.24$) into middle childhood ($a^2=.57$ and $.63$) due to new genetic influences. In addition, the authors reported an increase in nonshared environmental effects on adolescent language ($e^2 = .13$ and $.22$) compared to early ($e^2 = .02$) and middle childhood ($e^2=.06$ and $.05$). One particularly interesting feature from this study was the diversity of measures used to assess child language across ages: early childhood focused on a parent report measure of combined vocabulary and syntax; middle childhood included both a standardized vocabulary subtest and a curriculum-based teacher assessment of “speaking and listening”; and adolescence included a web-based assessment of receptive vocabulary, syntax, and pragmatics derived from standardized language subtests and a teacher report measure similar to middle childhood. Consequently, it is difficult to disentangle whether results reflect shifting etiological effects across time points or differences in the underlying language constructs being assessed.

The Western Reserve Reading and Math Project (WRRMP; Petrill, Deater-Deckard, Thompson, DeThorne, & Schatschneider, 2006), a longitudinal twin study of reading development and related abilities, is unique in that it includes language sample data collected during annual home visits in addition to standardized language assessments. Using this sample, DeThorne et al. (2008) conducted a multivariate genetic analysis of language skills that included 380 twins at first grade. Language measures loaded on two latent factors: a Conversational factor that included language sample measures such as mean length of utterance (MLU), number of total words (NTW), and number of different words (NDW), and a Formal factor that included two standardized vocabulary tests. Heritability for the Conversational and Formal factors was .70 and .37 respectively, with a genetic correlation of .37 between the two factors.

A follow-up study by DeThorne, Harlaar, Petrill, and Deater-Deckard (2012) examined the longitudinal stability of genetic and environmental influences on children's language sample measures, again using a Conversational language factor, across first and second grades. The analyses revealed that 62% of the variance in children's Conversational language at first-grade was due to genetic effects. At second grade, the heritability of the same Conversational factor dropped to 34% and overlapped entirely with the genetic effects from one year prior. In contrast to the stability in genetic effects, the nonshared environmental effects, estimated at 38% and 66% at first and second grade respectively, did not overlap across the two time points.

The present work represents an extension of prior WRRP analyses (i.e., DeThorne et al., 2008; 2012) by examining the extent and stability of etiological influences on children's language development, using both language sample and standardized measures, across the critical developmental window of early adolescence. Specifically we asked the following research questions:

1. To what extent do genetic and environmental factors contribute to individual differences in children's language skills, measured both by language sampling and standardized tests, during early adolescence?
2. To what extent are genetic and environmental influences on language stable across the three time points?

Method

Participants

Participants from the present study came from the Western Reserve Reading and Math Project (WRRMP; Petrill et al., 2006). Same-sex twins from Ohio were recruited during kindergarten/first grade through media advertisements, school nominations, Ohio state birth records, and Mothers of Twins clubs. The cohort was systematically followed through a series of annual home visits from kindergarten until at least 5th grade. The present analyses focused on data from the fifth, sixth, and seventh home visits², which centered on grades third, fourth, and fifth respectively. Each annual home visit included a 2.5 hour assessment protocol of standardized measures and parent questionnaires. At each home visit, twins were assessed simultaneously, each by a separate examiner.

Demographic data, such as racial/ethnic background and parent education, were taken from questionnaires administered at the time families entered the study. Across the three time points, the sample was 91-93% White, 57-58% female, and included 21-42% MZ (vs. DZ) twins. According to parent reports of highest educational level attained, 100% of the respondents had completed high school and 92% of them had pursued post-secondary education: 32% of parents had completed a four-year degree and 22% had completed graduate or professional school.

² Given that an extra semi-annual home visit was conducted between the third and fourth annual home visits, the fourth *annual* home visit is referred to as the fifth home visit (HV5) in the present paper.

Zygoty was determined through DNA testing from buccal swabs or a questionnaire of twin similarity reported to be 95% accurate (cf. Goldsmith, 1991). Information regarding hearing health, expressive language development, and speech-language pathology services was collected from a primary caregiver at entry into the study through the Speech-Language Survey (DeThorne et al., 2006).

Twins were selected from WRRP on the basis of having at least one type of language data (either language samples or standardized language test scores) at each home visit, as well as complete data on age, biological sex, and zygosity; no opposite-sex pairs were included in the study. Five individuals were subsequently excluded due to a history of persistent hearing difficulties, as reported by parents on the Speech-Language Survey. This process identified 522 participants at Home Visit 5 (HV5), of whom 95% had both language samples and standardized language test scores, 498 participants at Home Visit 6 (HV6), of whom 87% had both language sample and standardized language test scores, and 504 participants at Home Visit 7 (HV7), of whom 78% had both language sample and standardized language test scores. A total of 436 participants completed all three home visits. The mean age of twins was 9.83 years ($SD = .99$) at HV5, 10.88 years ($SD = 1.02$) at HV6, and 12.21 years ($SD = 1.21$) at HV7. Based on results from the Speech-Language Survey, between 7-8% of the sample was receiving speech-language pathology services when they entered the study.

Procedures

Language data collected during each annual home visit included standardized test scores from three Clinical Evaluation of Language Fundamentals-4th ed. subtests (Semel, Wiig, & Secord, 2003) and the Test of Narrative Language (TNL; Gillam & Pearson, 2004). In addition, the narratives produced during the TNL were supplemented by one additional picture plate taken

from stimulus manual one of the Test of Language Competence-Expanded Edition (TLC-E; Wiig & Secord, 1989). The supplemental picture plate, which included a visual scene of three children, one of whom appeared to have gotten off of her bike because of a branch stuck through her spokes, was added to the protocol to lengthen the narrative samples. Similar to the final task in the TNL, students were asked to look carefully at the photo and tell a related story. This picture elicitation task paired with the 3 expressive tasks from the TNL formed the narrative language sample. The entire narrative sample was digitally recorded and transcribed according to conventions associated with Systematic Analysis of Language Transcripts (SALT; Miller, Iglesias, & Nockerts, 2004). To prevent the inflation of utterance length based on multiple conjoining conjunctions, independent clauses joined by common conjunctions (i.e., *and*, *but*, *or*) were segmented into separate utterances forming what has been referred to as *communication units*, or C-units (Loban, 1976; Nippold, 1998). In addition, repeated or reformulated units were removed from the linguistic analysis through parenthetical mazes.

Transcription was completed by research assistants within the Child Language and Literacy Laboratory at the University of Illinois. Twins within a pair were transcribed by different research assistants so as not to procedurally inflate twin similarity. In addition, assistants were kept naïve as to the zygosity of the twins they were transcribing. All assistants were trained through iterative review of the SALT tutorial and lab manual until transcription of practice samples resulted in 85% point-by-point agreement with an experienced transcriber on both utterance boundaries and individual morphemes. In addition, approximately every 15th sample transcribed was subjected to a transcription reliability check in order to monitor and remedy potential drift in practiced conventions across transcribers. Based on this process, the point-by-point agreement across HV5 samples ($n=44$) ranged from .67-1.00 (mean = .92) for

utterance boundaries and .92-.98 for morphemes (mean=.96). Similarly, reliability for HV6 samples ($n=31$) ranged from .62-1.00 (mean = .93) for utterance boundaries and .72-.99 for morphemes (mean = .95). Finally, HV7 samples ($n=33$) ranged in agreement from .84-1.00 for utterance boundaries (mean = .94) and .89-.99 for morphemes (mean = .97).

The transcription process resulted in language samples with an average length of 59 C-units at HV5 ($SD = 31$, range = 17 to 297), 62 at HV6 ($SD = 25$, range = 14-243) and 60 at HV7 ($SD = 23$, range = 28 to 236). Our elicitation procedures (i.e., story retell and single-picture elicitation method) and resulting sample lengths are well in line with prior published language sample analyses of narrative samples (e.g., Miller et al., 2006; Nippold, Hesketh, Duthie, & Mansfield, 2005; Scott & Windsor, 2000; Tilstra & McMaster, 2007). A study by Heilmann et al. (2010) explicitly examined the reliability of productive language measures from both conversational and narrative samples. Most relevant to the present study, they reported Cronbach's alpha values of .74, .93, and .92 for MLU, NDW, and Words Per Minute (a measure comparable to NTW) on narrative samples of comparable length to our study in children ages 6;0-13;3 (pp. 398).

Standardized Test Scores.

Test of Narrative Language (TNL). The present analysis included the Oral Narration score from the TNL, which is based on 3 expressive tasks: retelling a story about a trip to a fast food restaurant without any picture prompts, telling a narrative related to a sequence of five drawings of a boy late for school, and creating a story based on a single pictured scene in which two children appear to have stumbled across aliens landing in a park. Consistent with scoring procedures reported in the manual, results across the three expressive tasks were summed into a single raw score for Oral Narration with a maximum possibility of 90 points. Designed for

children age 5;0 to 11;11, the internal consistency (coefficient alpha) for the Oral Narration score reported in the test manual ranged from .85 to .90 for children between the ages of 8 and 11 (p. 44). In addition, test-retest reliability for the Oral Narration score from a relatively small sample of children age 5-10 years was reported as .80 (uncorrected).

Clinical Evaluation of Language Fundamentals (4th ed.) subtests.

Recalling sentences. Considered an assessment of expressive language and memory, this subtest involved repetition of spoken sentences such as “The rabbit was not put in the cage by the girl.” and “Does anyone know who the new teacher is?” Raw scores could range from 0-98 based on the number of completed items and the number of errors. Errors included omissions, repetitions, additions, transpositions, and substitutions. Subtest reliability reported in the manual ranged from .87 to .92 for test-retest and from .86 to .91 for internal consistency (alpha) across relevant age ranges (9-15 years).

Understanding Spoken Paragraphs. Recognized as a measure of receptive language, this subtest involved asking children to answer questions about three short narratives read aloud (4-9 sentences), which varied based on age. There were a total of 15 questions focused on main ideas, details, sequences, inferences, and predications related to the spoken narratives. Raw scores could range from 0-15 regardless of child age. Subtest reliability reported in the manual ranged from .51 to .87 for test-retest and from .54 to .68 for internal consistency (alpha) across relevant age ranges (9-15 years).

Word Classes 2 Receptive and Expressive. Each item from the *Word Classes* subtest involved a receptive and expression portion, and required the child to listen to a set of four words read aloud. For the receptive portion of the task, the child had to state *which* two items from each set were related. For the expressive portion, the child had to state *how* the two items they

selected were related. Raw scores for both the receptive and expressive portions could range from 0-24, for a range of 0-48 when summed together. Reliability data for the summed scores (receptive + expressive) ranged from .83 to .90 for test-retest and from .87 to .91 for internal consistency (alpha) across relevant age ranges (9-15 years).

Language sample measures.

Number of total words (NTW). The number of total words, a frequency count of all root word tokens within the first 30 complete and intelligible C-units, was derived as a general measure of word productivity. To simultaneously minimize the potential influence of volubility while also maximizing the proportion of available data, NTW was derived from the first thirty complete and intelligible C-units in each sample³. A prior study from WRRP (i.e. DeThorne, Deater-Deckard, Mahurin-Smith, Coletto, & Petrill, 2011) directly examined volubility as a potential confound for measures of NTW and NDW, and found that both measures did *not* correlate with measures of child temperament, including surgency/extraversion, when derived from a standardized number of utterances.

Number of different words (NDW). To provide a measure of vocabulary diversity, the number of different word tokens used by each child was also derived from the first 30 complete and intelligible C-units within each sample (cf. DeThorne et al., 2011; Hutchins, Brannick, Bryant, & Silliman, 2005). We considered the alternative of deriving NDW from 100 tokens, consistent with Scott and Windsor (2000); however this measure did not lead to many additional cases and actually tended to rely on shorter samples (i.e., samples of 30 C-units had an average of around 250 tokens as compared to 100 tokens).

³ Calculating NTW based on 100-utterances, consistent with our prior work (e.g., DeThorne et al., 2012), would have resulted in too few usable cases given that many of the narrative samples were relatively short (mean ranging from 59-60 C-units across the three home visits)

Mean Length of utterance (MLU). The average length of child utterances in morphemes was derived from all complete and intelligible C-units as a measure of children's productive abilities, including morphosyntactic development. Consistent with conventions (Miller et al., 2005), bound inflectional morphemes were counted as separate morphemes but derivational morphemes were not.

Although evidence regarding the reliability of productive language samples measures has been mixed, especially from shorter samples (e.g., Gavin & Giles, 1996; Heilmann, Nockerts, & Miller, 2010; Tilstra & McMaster, 2007), such measures have been associated with evidence of validity through developmental change with age up through age 13 (Miller, Freiberg, Holland, & Reeves, 1992; Miller et al., 2005; Redmond, Rice, & Hoffman, 2006), correlation with other measures of linguistic complexity (DeThorne et al., 2008; Nippold, 2009; Nippold et al., 2014; Rice et al., 2006; Ukrainetz & Blomquist, 2002), and differentiation of clinical groups (e.g., Condouris, Meyer, & Tager-Flusberg, 2003; Heilmann, Miller, & Nockerts, 2010; Scott & Windsor, 2000; Watkins, Kelly, Harbers, & Hollis, 1995).

Analyses

The primary goal of this study was to examine the stability of etiological effects on language development during early adolescence. This was addressed through structural equation modeling of correlated latent factors derived from the language sample measures (referred to here as the Productive language factor)⁴ and from the standardized tests (Formal language factor). At each home visit, the Productive language factor was indexed by MLU, NTW and NDW, whereas the Formal language factor was indexed by TNL and the three CELF subtests.

⁴ DeThorne et al. (2008) used the term Conversational language for the factor created from language sample measures. The term Productive language is used here instead (consistent with Scott & Windsor, 2000) given that language samples were taken from narrative tasks rather than conversation.

These factors are consistent with the conceptual framework of prior studies (DeThorne et al., 2008, 2012; Heilmann et al., 2010; Miller et al., 2006; Scott & Windsor, 2000) and the phenotypic correlation data from the present study. Latent factors represent the shared variance amongst measures, independent of measure-specific variance and measure-specific error. The models are depicted in Figure 1 for Productive language ability and Figure 2 for Formal language test performance.

Importantly, the models shown in Figures 1 and 2 include genetic (A), shared environmental (C), and nonshared environmental (E) latent factors that decompose the variance *within* and covariance *between* the latent Productive language and Formal language factors. These parameters are arranged in a Cholesky decomposition formation, meaning that the first set of A, C, and E factors (denoted as A_1 , C_1 , and E_1) account for the variance in Productive language and Formal language test performance at home visit 5 as well the covariance, or stability, of Productive (Figure 1) or Formal (Figure 2) language ability across home visits 5, 6, and 7. One advantage of using latent factors rather than individual measures in the Cholesky modeling is that nonshared environmental effects (e^2) reflect individualized effects that are shared across all variables in the factor, thereby removing any measure-specific error (though not measurement error that is correlated across measures, reflecting, for example, common method variance; Podsakoff et al., 2003). The second set of A, C, and E factors (A_2 , C_2 , and E_2) represent genetic and environmental effects that contribute to stability between Productive or Formal language ability at home visits 6 and 7 over and above genetic and environment effects that influence Productive language ability at home visit 5. The third set of A, C, and E factors (A_3 , C_3 , and E_3) account for residual variance in Productive language ability and Formal language test performance at Home Visit 7 that is independent of both Home Visits 5 and 6; that is, “new”

genetic and environmental influences that emerge later in development and are not associated with language ability at earlier assessments. In addition to examining genetic and environmental influences on latent Productive and Formal language factors, our model included parameters to estimate measure-specific (residual) A, C, and E effects on each manifest variable (i.e., MLU, NTW and NDW in the model for Productive language, and TNL and the three CELF subtests for Formal language). The measure-specific E effects incorporate measure-specific error variance. The total effects of A, C, and E on each of the measures can be estimated as the sum of the measure-specific A, C, and E effects and the effects that are shared with the other variables loading onto the common latent factor, weighted for that variable's loading on the latent factor.

We used these models to estimate the relative contributions of A, C, and E to the phenotypic variance in Productive language and Formal language, which allowed us to address our first research question regarding the extent of genetic and environmental effects on children's language skills. Similar to DeThorne et al. (2012), we then decomposed the sources of the temporal stability of Productive and Formal language across the three home visits into: (1) A, C, and E effects that are stable across all three home visits, (2) A, C, and E effects that are stable across the last two visits only, and (3) A, C, and E effects that are specific to the last visit. This analysis allows us to address our second research question regarding the extent of genetic and environmental stability across the three time points. A developmental stability hypothesis predicts that genetic influences on language originate in the first latent genetic factor (A_1), with no significant subsequent genetic influences (i.e., path coefficients from A_2 and A_3 are not significantly different from zero). In contrast, a developmental change hypothesis predicts that new genetic variation will emerge at later home visits to affect language (i.e., path coefficients

from A_2 and/or A_3 are significantly different from zero). Parallel hypotheses for shared environmental effects and nonshared environmental effects can also be tested.

All models were estimated in the *Mx* structural equation modeling package (Neale, Boker, Xie & Maes, 2006) using full-information maximum likelihood (FIML) to handle missing data. By fitting the model to twin data, we can estimate genetic and environmental effect sizes and 95% confidence intervals for the parameter estimates. Prior to analyses, language scores were standardized on the whole sample to a mean of zero and a *SD* of 1. Outlying scores, defined as -3 or $+3$ *SD* from the mean, were removed. For the twin analyses, standardized residuals correcting for age and sex were used because the age of twins is perfectly correlated across pairs, which means that, unless corrected, variation within each age group at the time of testing would contribute to the correlation between twins and be misrepresented as shared environmental influence (McGue & Bouchard, 1984). The same applies to the sex of the twins, because monozygotic twins are always the same sex.

Results

Preliminary analyses

We considered two forms of evidence regarding the validity of our language sample measures. First, we examined potential group differences in each language sample measure as a function of whether or not caregivers reported a history of expressive language difficulties at enrollment in the study. A consistent trend for higher productive language scores for children without a history of reported expressive language difficulties emerged from all measures across home visits, though differences reached statistical significance in only three instances. In terms of effect size, Cohen's *D* ranged from $-.14$ in the case of MLU at HV6 to $-.71$ for NDW at HV5.

Though the parent report measure offers a positive indicator of social validity, the measure was taken approximately 3 years prior to the language sample measures being studied here. Consequently, a second form of validity evidence focused on children's *concurrent* performance using scores from the Recalling Sentences subtest from the CELF. Given prior evidence that sentence repetition serves as a relatively sensitive marker of language impairment (e.g., Conti-Ramsden, Botting, & Faragher, 2001; Meir, Walters, & Armon-Lotem, 2015; Stokes, Wong, Fletcher, & Leonard, 2006), we selected participants from each HV who scored at least one standard deviation above or below the mean on this subtest at each time point and compared the two groups on their productive language measures at the same home visit. Significant group differences emerged across all nine comparisons (3 productive language sample measures x 3 HVs), with Cohen's *D* varying from -.92 to -1.08 at HV5, from -.47 to -.70 at HV6, and from -.90 to -.96 at HV7. Although neither form of validity evidence is conclusive, together they provide some support for the reliability and validity of our productive language samples measures.

Descriptive statistics for each dependent measure, presented as raw scores, are summarized in Table 1. Results from ANOVAs revealed no significant mean or variance differences by zygosity or by classification of twins as twin 1 or twin 2 (reflecting birth order). Cross-measure phenotypic correlations within each home visit are shown in Table 2. There were high correlations among the language sample measures (.75-.90 among MLU, NTW, & NDW), moderate to high correlations among the standardized language measures (.31-.67) and moderate correlations across the language sample measures and the standardized language measures (.21-.46).

We applied a phenotypic confirmatory factor analysis (that did not include the effects of A, C, and E factors) to the language sample measures and standardized test scores to further test our conceptual distinction between these two types of assessment. This analysis supported the distinction between a Productive factor for the language sample measures (MLU, NTW, & NDW) and a Formal factor for the standardized scores (TNL, CRLF-RW, CELF-USP, & CELF-WC). The loadings of the language sample measures on the latent Productive factor at each home visit were substantial ($> .80$ at each home visit), whereas loadings of the standardized language measures on the Formal factor were moderate to large ($.30$ -. $.44$ for TNL, $.88$ -. $.91$ for CELF-RS, $.41$ -. $.56$ for CELF-USP, and $.65$ -. $.73$ for CELF-WC)⁵. As previously mentioned, the latent Productive language and Formal language factors represent the shared variance among the measures, independent of measure-specific variance and measure-specific error.

In terms of longitudinal phenotypic analysis, the correlations between the Productive latent factors at HV5 and HV6 and between HV6 and HV7 were moderate (both correlations $r = .31$), and about double the magnitude of the phenotypic correlation between Productive language at HV5 and HV7 ($r = .15$). In contrast, the phenotypic correlations for the Formal language factor across the home visits (HV5 and HV6, HV6 and HV7, HV5 and HV7) were unity ($r=1.00$), indicating that the common variance captured by the latent factors representing standardized language were perfectly stable.

Q1: To what extent do genetic and environmental factors contribute to individual differences in children's language skills, measured both by language sampling and standardized tests, during early adolescence?

⁵ An alternative model in which TNL was set to load on the latent Productive language factor, rather than Formal language factor, resulted in a poorer model fit (details available in the online supplementary materials; Table S1), and was therefore not considered further.

An initial impression of the extent to which genetic and environmental factors contribute to individual differences in children's language skills can be gleaned from intraclass twin correlations (ICC), which are summarized in Table 3 for the two language factors at each home visit. For the Productive language factor, correlations were small for both MZ and DZ twin pairs and reached statistical significance only at HV6. This pattern suggests that nonshared environmental influences accounted for the majority of the phenotypic variance in the language sample measures. In contrast, the Formal language factor demonstrated a consistent trend toward higher similarity between MZ pairs relative to DZ pairs, with large significant correlations for MZ pairs that were almost twice the magnitude of the DZ twin correlations. This pattern is suggestive of high heritability.

Estimates of the proportion of variance in the latent Productive and Formal language factors that are due to genetic and environmental effects (derived from the genetically-informative Cholesky decomposition models shown in Figures 1 and 2, respectively) are largely aligned with the intraclass correlations. As shown in Table 4, variance in the Productive language factor at each age was primarily due to nonshared environmental influences, which ranged from 90% of the variance at HV5 to 55% at HV7. Genetic (i.e., heritable) and shared environmental influences were small and non-significant. In contrast, variance in the Formal language factor at each age was primarily due to genetic influences, which ranged from 82% at HV5 to 86% at HV7. Both shared and non-shared environmental influences were small. Estimates for the non-shared environment, but not the shared environment, were significantly different from zero, ranging in effect size from 5-6% of the variance for the Formal language factor. Finally, measure-specific E effects on the manifest variables, which include measurement-specific error variance, were generally significantly different from zero in both the

Productive language and Formal language models. In contrast, measure-specific A and C effects on each manifest variable were not significant for any of the measures. Estimates for these measure-specific effects, as well as the total genetic, shared and nonshared environmental influences on each of the measured variables, are shown in the online supplementary materials (Tables S2 and S3 for Productive language and Formal language, respectively).

Q2: To what extent are genetic and environmental influences on language stable across the time points?

The Cholesky decomposition model also provided information on the stability of genetic and environmental factors across the three home visits, and thus allowed us to address our second research question. The full set of standardized parameter estimates from the models for both Productive and Formal language are included in the online supplementary materials. Here, we focus on the squared standardized parameter estimates, which provide information on the extent to which the variance in Productive and Formal language at each home visit is due to additive genetic, shared environmental and nonshared environmental influences that are shared (i.e., stable) across home visits. Table 5 shows the squared, standardized estimates for Productive language. For reference, the proportions of variance in the latent Productive language factor due to additive genetic, shared environmental, and nonshared environmental influences (from Table 4) are shown in bold.

The primary variance component of interest is nonshared environmental factors, because the 95% confidence intervals from Table 4 indicate that neither additive genetic nor shared environmental factors made a significant contribution to variance in Productive language at any home visit. Nonshared environmental effects common to all three factors (E_1 in Figure 1) accounted for all of the nonshared environmental variance at HV5 ($.90/.90 = 1$, where the

denominator, .90, refers to the total nonshared environmental variance in Productive language at the home visit of interest – in this case, HV5). It also accounted for around 5% of the nonshared environmental variance at HV6 ($.04/.76 = .05$), and 2% of the nonshared variance at HV7 ($.01/.55 = .02$). Nonshared environmental factors specific to Home Visits 6 and 7 (E_2 in Figure 1) accounted for the majority, and the remaining (95%) nonshared environmental variance in HV6 ($.74/.76 = .95$), and 2% of the nonshared variance at HV7 ($.01/.55 = .02$). Finally, nonshared environmental effects specific to HV7 (E_3 in Figure 1) accounted for the remaining (96%) nonshared environmental variance in HV6 ($.53/.55 = .96$). In sum, there was evidence for nonshared environmental stability in Productive language, but overall effect sizes were small, with most nonshared environmental effects being specific (i.e., new) to each home visit.

Parallel evaluations can be made for Formal language. Table 6 shows the squared, standardized estimates of the Cholesky decomposition model of Formal language. Because shared environmental effects were not significantly different from zero (Table 4), the primary variance components of interest are genetic and nonshared environmental factors. Genetic effects common to all three factors (represented by A_1 in Figure 2) accounted for all of the genetic variance at HV5 (.82). It also accounted for 100% of the genetic variance at HV6 ($.83/.83 = 1$) and at HV7 ($.86/.86 = 1$). Genetic factors specific to Home Visits 6 and 7 (represented by A_2) and genetic factors that influenced only Home Visit 7, independent of Home Visits 5 and 6 (represented by A_3), had zero effect on Formal language at later home visits. The same pattern emerged for nonshared environmental factors, even though the proportion of variance in Formal language due to nonshared environmental factors was much smaller than the proportion of variance due to genetic effects. Nonshared environmental effects common to all three factors (represented by E_1) accounted for all of the nonshared environmental variance at Home Visits 5

(.05), 6 (.05), and 7 (.06). Overall, these estimates point to a pattern of strong developmental stability in genetic and environmental effects on Formal language across early adolescence.

Fit statistics for nested submodels for both Productive and Formal language are included in the Supplementary Online Materials (Tables S4 and S5, respectively), to facilitate comparisons between the full models and models in which we dropped non-significant parameters. These fit statistics indicate that the fit of the full models was generally comparable or better than the nested submodels.

Discussion

The present study of etiological influences on child language during early adolescence found strikingly different patterns of effects depending on whether or not language was measured through standardized assessments or less structured productive tasks. Specifically, the Formal language factor demonstrated a strong and consistent pattern of genetic effects at each time point ($a^2 = .82, .83, .86$), with significant but small nonshared effects at each time point as well ($e^2 = .05, .05, .06$). In contrast, the Productive language factor demonstrated a strong pattern of nonshared environmental effects at each visit ($e^2 = .90, .76, .55$) with no significant genetic or shared environmental effects. In terms of longitudinal stability within each factor, we found complete overlap in the genetic and nonshared environmental effects on the Formal language factor across visits, as well as significant but limited stability in nonshared environmental effects on the Productive language factor for adjacent time points. The following discussion will center on: a) interpreting the evidence for differing etiologies of the two language factors, b) integrating the information about etiological stability with prior research, c) noting study limitations, and d) highlighting implications for clinical practice and future research.

Interpreting the differing etiologies between the two language factors

One potential interpretation of the etiological differences between Productive and Formal language factors, particularly in light of the strong nonshared environmental effects on the Productive language factor, is that the narrative language samples did not provide a reliable form of measurement. This is particularly important given the fact that latent factors do not control for forms of measurement error that might be shared across measures (e.g., rapport with examiner, interest in the task). Despite this reasonable concern, support for the reliability and validity of our productive language measures comes from at least three sources. First, our elicitation task, transcription procedures, and sample length are in accord with prior published procedures (e.g., Nippold et al., 2005; Miller et al., 2006; Scott & Windsor, 2000; Tilstra & McMaster, 2007). Second, we found relatively consistent validity evidence for our language sample measures through group comparisons. Specifically, whether grouped based on parents' report of expressive language history or concurrent performance on the CELF-RS subtest, significant group differences in the productive language measures consistently emerged in the anticipated direction. Finally, it seems important to note that the significant stability in nonshared effects across time points, albeit small in effect size, does lend additional support to the validity of our productive language measures. It is interesting to note that even if one attributes the nonshared environmental influences on the Productive language factor to measurement error/poor reliability, this result is still of critical importance for interpreting other studies that have relied on similar language sampling procedures.

If one accepts support for at least partial validity of the productive language measures, then the etiological differences across the Productive and Formal language are intriguing, suggesting either that an individual's language skills can differ greatly across contexts or that language assessments are influenced by traits other than linguistic prowess. In fact, the two

possibilities are likely intertwined. Situated accounts of communication have highlighted that language use is always embedded within specified contexts (cf. Hengst, 2015), and prior literature has documented the impact of factors such as attention, motivation, compliance, frustration tolerance, persistence, anxiety, and cultural background on test-taking performance (Allan, 1992; Dreisbach & Keogh, 1982; Erickson, 1972; Fleege, Charlesworth, Burts, & Hart, 1992; Peña, Iglesias, & Lidz, 2001; Speltz, DeKlyen, Calderon, Greenberg, & Fisher, 1999). Similarly, productive measures such as MLU, NTW, and NDW have been noted to differ based on such factors as motivation, genre, modality, place, and partner (e.g., Bornstein, Haynes, Painter, & Genevro, 2000; Fields & Ashmore, 1980; Nippold et al., 2009, 2014; Scott & Windsor, 2000).

In terms of our Formal versus Productive language factors, a couple of potential differences come to mind, specifically language modality (expressive *vs.* receptive) and degree of on-demand constraints. Specifically, the Formal factor included two subtests that explicitly measured receptive language abilities, whereas the Productive factor, as the name suggests, focused on expressive abilities. In addition, the measures in the Formal language factor, particularly the CELF subtests, are fairly constrained and adult-directed, meaning children have less opportunity to showcase their linguistic skills aside from the specified targets. Many of us have been in situations where we want to credit a child for her test response because it feels clever or creative even if it isn't the 'expected' right answer. In contrast, narrative samples offer children more latitude for linguistic flexibility and creativity. For example, one child colorfully used the adjective "seafoam green" in her narrative description of a boy's shirt; such vocabulary is unlikely to be directly queried in a standardized assessment.

In sum, perhaps it is not particularly surprising that the etiological influences on language outcomes would differ in part based on how it is being assessed. A number of studies have highlighted that language sample measures correlate more strongly with each other than with standardized tests, and that standardized tests across multiple domains tend to correlate with each other (cf. Condouris et al., 2003; DeThorne, Johnson, & Loeb, 2005; DeThorne & Watkins, 2006; Ukrainetz & Blomquist, 2002). Of course from a psychometric standpoint, influences other than the construct of interest are often chalked up to “error”; however, it is interesting to consider just how much the concept of error applies when findings are stable both within forms of assessment (e.g., productive language sample, standardized tests) and across measurement time points.

Integrating the information about etiological stability with prior research

Given the distinct and consistent etiological patterns associated with the Productive versus Formal language factors, it seems important to reconsider past behavioral genetic findings with direct consideration of form of measurement. Specifically, Haworth et al. (2010) reported linear increases in the heritability of general cognition from childhood to young adulthood using standardized IQ measures. Hayiou-Thomas et al. (2012) reported increased heritability in child language from early to middle childhood but not from middle childhood into adolescence. Of particular interest, the early childhood measure in Hayiou-Thomas et al. (2012) was collected through parent report, whereas the middle childhood and adolescent time points included direct standardized assessment and teacher curricular assessments, consequently making it difficult to differentiate age from form of assessment effects. Regardless, Hayiou-Thomas et al. (2012) reported similar etiological findings across both direct assessments and teaching ratings in mid childhood and adolescence. To the extent that these findings are akin to our Formal and

Productive language factors respectively, then the two studies' findings would appear to be discrepant for the productive/report measure (yet very similar for the formal/direct measures).

Compared with our prior work, the present study found higher heritability for the Formal factor during early adolescence (ranging from .82 to .84 across time points) when compared to the .45 reported for the same general participant pool at earlier ages (DeThorne et al., 2008, p.8). In addition, findings suggest a trend toward decreasing heritability in productive language sample measures when current estimates, obtained during early adolescence, are compared to first- and second-grade time points (DeThorne et al., 2012). Taken together, findings across studies support the proposition that the heritability of child language development increases across development *as assessed through standardized measures* but perhaps not for language measures taken from other forms of assessment, such as discourse-based language samples.

Study limitations

The contributions of the present study should be interpreted in light of limitations associated with the design and the measures employed. In terms of design, behavior genetic studies rely on large samples, and the present study was underpowered to detect small effect sizes. Consequently, our data should not be used to suggest that there are no shared environmental effects on individual differences in early adolescent language development, nor that genetic effects no longer contribute to differences in child language sample measures at early adolescence. Second, behavioral genetic designs by nature are intended to explain etiological contributions to variance within a given population. The restricted composition of our sample in terms of variation in race/ethnicity and parent education limits our examination of variance associated with such factors. Finally, behavioral genetic analyses are not well positioned to examine the complexities of epigenetic effects, specifically environmental

influences on genetic expression (cf. Kraft & DeThorne, 2014; Rogers, Nulty, Aparicio Betancourt, & DeThorne, 2015).

Highlighting implications for clinical practice and future research

In terms of implications for research, studies of child language should be sure to incorporate multiple forms of assessment, both for diagnostic purposes and in assessment of outcome variables. This recommendation has been stressed for clinical purposes, but less so for research. In addition, our significant findings for nonshared environmental effects on both standardized and productive language measures suggest that additional study of nonshared environmental factors is warranted. The significance of nonshared environmental effects tells us that *individualized* experiences shape language development, which to some extent is difficult to reconcile with the field's heavy focus on caregiver interaction style and amount of linguistic input (cf. Rogers et al., 2015). Given that caregivers are generally the same within twin pairs, such variables could not be contributing in isolation to the significant nonshared environmental effects observed here. Candidates for potential nonshared environmental effects include such factors as diet, brain injury, peer affiliations, classroom placement, sleep patterns, drug/toxin exposure, and stress. In considering such issues, it is not difficult to begin to see how complex and individualized the web of causal influences on human development truly is.

In regard to clinical applications, the significance of genetic and nonshared environmental effects on adolescent language development supports the need to approach child language development much more individualistically, recognizing that each child brings both unique capacities and experiences to the language-learning task. Consequently, there is not likely to be a “one-size-fits-all” approach or strategy for every child. Although many clinicians would likely agree with this assertion in theory, clinical and educational practice often revolve around

packaged programs and curricula that may not directly acknowledge individual differences. Similarly, reductionist experimental group paradigms often focus on mean differences across groups rather than the complexity of individual differences, thereby cultivating the impression that one approach is preferred for all. In closing, findings of significant genetic and nonshared environmental influences on individual differences in early adolescent language development call us to think explicitly about the complexity of *individualized* language learning trajectories, a challenge well-suited, one would hope, to a discipline such as communication sciences and disorders.

Acknowledgments

This project was supported by multiple grants from the National Institute of Child Health and Human Development (HD38075, HD46167, HD050307). In addition to our appreciation of all participating families, we want to extend thanks to other WRRMP co-investigators: Kirby Deater-Deckard, Lee Anne Thompson, & Christopher Schatschneider. Finally, thanks to all participating members of the Child Language & Literacy Lab, particularly Karissa Nulty and Clare Rogers for assistance with data management and analyses.

Figure Legends

Figure 1. Cholesky decomposition model of productive language measured at three home visits (HV). Latent additive genetic (A), shared environmental (C) and nonshared environmental (E) factors account for variance in the latent productive language factor at each home visit (A, C, and E factors shown at the top of the figure) as well as residual variance in the individual language measures (A, C, and E factors shown at the bottom of the figure). Individual productive language measures are MLU (Mean Length of Utterances), NDW (Number of Different Words), and NTW (Number of Total Words).

Figure 2. Cholesky decomposition model of formal language measured at three home visits (HV). Latent additive genetic (A), shared environmental (C) and nonshared environmental (E) factors account for variance in the latent formal language factor at each home visit (A, C, and E factors shown at the top of the figure) as well as residual variance in the individual language measures (A, C, and E factors shown at the bottom of the figure). Individual formal language measures are TNL (Test of Narrative Language), RS (CELF-Recalling Sentences), USP (CELF-Understanding Spoken Paragraphs), and WC (CELF-Word Classes).

References

- Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers. *Language Testing*, 9, 101–122.
- American Psychological Association. (2002). *Developing adolescents: A reference for professionals*. Washington, DC. Retrieved from <http://www.apa.org/pi/families/resources/develop.pdf>
- Blakemore, S. J. (2012). Development of the social brain in adolescence. *Journal of the Royal Society of Medicine*, 105(3), 111-116. doi: 10.1258/jrsm.2011.110221
- Bornstein, M. H., Haynes, M., Painter, K. M., & Genevro, J. L. (2000). Child language with mother and with stranger at home and in the laboratory: A methodological study. *Journal of Child Language*, 27(02), 407–420. Retrieved from <http://journals.cambridge.org/action/displayJournal?jid=JCL>
- Condouris, K., Meyer, E., & Tager-Flusberg, H. (2003). The relationship between standardized measures of language and measures of spontaneous speech in children with autism, *American Journal of Speech-Language Pathology*, 12(3), 349-358.
- Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology & Psychiatry*, 42(6), 741-748.
- Dahl, R. E. (2004). Adolescent brain development: A period of vulnerabilities and opportunities. Keynote address. *Annals of the New York Academy of Sciences*, 1021(1), 1-22. doi: 10.1196/annals.1308.001
- DeThorne, L. S., & Hart, S. A. (2009). Use of the twin design to examine evocative gene-environment effects within a conversational context. *International Journal of Developmental Science*, 3(2), 175-194. doi: 10.3233/DEV-2009-3207

- DeThorne, L. S., & Watkins, R. V. (2006). Language abilities and nonverbal IQ in children with language impairment: Inconsistency across measures. *Clinical Linguistics & Phonetics*, 20(9), 641-658. doi: 10.1080/02699200500074313
- DeThorne, L. S., Hart, S. A., Petrill, S. A., Deater-Deckard, K., Thompson, L. A., Schatschneider, C., & Davison, M. D. (2006). Children's history of speech-language difficulties: Genetic influences and associations with reading-related measures. *Journal of Speech, Language, & Hearing Research*, 49(6), 1280-1293. doi: 10.1044/1092-4388(2006/092)
- DeThorne, L. S., Johnson, B. W., & Loeb, J. W. (2005). A closer look at MLU: What does it really measure?. *Clinical Linguistics & Phonetics*, 19(8), 635-648. doi: 10.1080/02699200410001716165
- DeThorne, L.S., Deater-Deckard, K., Mahurin-Smith, J., Coletto, M., & Petrill, S.A. (2011). Volubility as a mediator in the associations between conversational language measures and child temperament. *International Journal of Language and Communication Disorders*, 46(6), 700-713. doi: 10.1111/j.1460-6984.2011.00034.x
- DeThorne, L.S., Harlaar, N., Petrill, S.A., & Deater-Deckard, K. (2012). Longitudinal stability in genetic effects on children's conversational language productivity. *Journal of Speech, Language, & Hearing Research*, 55(3), 739-753. doi: 10.1044/1092-4388(2011/11-0014)
- DeThorne, L.S., Petrill, S.A., Hart, S.A., Channell, R.W., Campbell, R.J., Deater-Deckard, K., ... Vandenberg, D.J. (2008). Genetic effects on children's conversational language use. *Journal of Speech, Language, & Hearing Science*, 51(2), 423-435. doi: 10.1044/1092-4388(2008/031)
- Dreisbach, M., & Keogh, B. K. (1982). Testwiseness as a factor in readiness test performance of

- young Mexican-American children. *Journal of Educational Psychology*, 74(2), 224.
<http://dx.doi.org/10.1037/0022-0663.74.2.224>
- Erickson, M. E. (1972). Test sophistication: an important consideration. *Journal of Reading*, 16(2), 140–145.
- Fields, T. A., & Ashmore, L. L. (1980). Effect of elicitation variables on analysis of language samples for normal and language-disordered children. *Perceptual and Motor Skills*, 50(3), 911-919. doi: 10.2466/pms.1980.50.3.911
- Fleege, P. O., Charlesworth, R., Burts, D. C., & Hart, C. H. (1992). Stress begins in kindergarten: A look at behavior during standardized testing. *Journal of Research in Childhood Education*, 7(1), 20-26. doi: 10.1080/02568549209594836
- Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech, Language, and Hearing Research*, 39(6), 1258-1262. doi: 10.1044/jshr.3906.1258
- Gillam, R. B., & Pearson, N. A. (2004). *Test of narrative language: Examiner's manual*. Austin: TX: Pro-ed.
- Goldsmith, H. (1991). A zygoty questionnaire for young twins: A research note. *Behavior Genetics*, 21(3), 257-269. doi: 10.1007/BF01065819
- Haworth, C. M. A., Wright, M. J., Luciano, M., Martin, N. G., De Geus, E. J. C., Van Beijsterveldt, C. E. M., ... & Plomin, R. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. *Molecular Psychiatry*, 15(11), 1112-1120. doi: 10.1038/mp.2009.55
- Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R. (2012). The etiology of variation in language skills changes with development: A longitudinal twin study of language from 2 to 12

- years. *Developmental Science*, 15(2), 233-249. doi: 10.1111/j.1467-7687.2011.01119.x
- Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools*, 41(1), 84-95. doi: 10.1044/0161-1461(2009/08-0075)
- Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter?. *Language, Speech, and Hearing Services in Schools*, 41(4), 393-404. doi: 10.1044/0161-1461(2009/09-0023)
- Hengst, J.A. (2015). Distributed communication: Implications of cultural-historical activity theory (CHAT) for communication disorders. *Journal of Communication Disorders*, 57, 16-28. doi: 10.1016/j.jcomdis.2015.09.001
- Herman-Giddens, M. E., Slora, E. J., Wasserman, R. C., Bourdony, C. J., Bhapkar, M. V., Koch, G. G., & Hasemeier, C. M. (1997). Secondary sexual characteristics and menses in young girls seen in office practice: a study from the Pediatric Research in Office Settings network. *Pediatrics*, 99(4), 505-512. doi: 10.1542/peds.99.4.505
- Hoekstra, R. A., Bartels, M., & Boomsma, D. I. (2007). Longitudinal genetic study of verbal and nonverbal IQ from early childhood to young adulthood. *Learning and Individual Differences*, 17(2), 97-114. doi: 10.1016/j.lindif.2007.05.005
- Hopper, J. L. (2000). Why “common environmental effects” are so uncommon in the literature. In T. D. Spector, H. Snieder, & A. J. MacGregor (Eds.), *Advances in twin and sib-pair analysis* (pp. 151–165). London, UK: Oxford University Press.
- Hutchins, T. L., Brannick, M., Bryant, J. B., & Silliman, E. R. (2005). Methods for controlling amount of talk: Difficulties, considerations and recommendations. *First Language*, 25(3), 347–363. doi: 10.1177/0142723705056376

- Huttenlocher, P. (1994). Synaptogenesis in human cerebral cortex. In G. Dawson, & K. W. Fischer (Eds). *Human behavior and the developing brain*. (pp. 137–152). New York: The Guilford Press.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology*, 21(1), 60-99. doi: 10.1016/0010-0285(89)90003-0
- Kraft, S. J., & DeThorne, L. S. (2014). The brave new world of epigenetics: Embracing complexity in the study of speech and language disorders. *Current Developmental Disorders Reports*, 1(3), 207-214. doi: 10.1007/s40474-014-0024-4
- Loban, W. (1976). *Language development: Kindergarten through grade twelve* (Research Report No. 18). Urbana, IL: National Council of Teachers of English.
- Marshall, W. A., & Tanner, J. M. (1970). Variations in the pattern of pubertal changes in boys. *Archives of disease in childhood*, 45(239), 13-23. doi: 10.1136/adc.45.239.13
- Marshall, W. A., & Tanner, J. M. (1969). Variations in pattern of pubertal changes in girls. *Archives of disease in childhood*, 44(235), 291.
- McGue, M., & Bouchard Jr, T. J. (1984). Adjustment of twin data for the effects of age and sex. *Behavior Genetics*, 14(4), 325-343. doi: 10.1007/BF01080045
- Meir, N., Walters, J., & Armon-Lotem, S. (2015). Disentangling SLI and bilingualism using sentence repetition tasks: the impact of L1 and L2 properties. *International Journal of Bilingualism*, doi: 10.1177/1367006915609240
- Miller, J. F., Freiberg, C., Holland, M. B., & Reeves, M. A. (1992). Implementing computerized language sample analysis in the public school. *Topics in Language Disorders*, 12(2), 69-82. doi: 10.1097/00011363-199202000-00008

- Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice*, 21(1), 30-43. doi: 10.1111/j.1540-5826.2006.00205.x
- Miller, J.F., Iglesias, A., & Nockerts, A. (2004). *SALT: Systematic analysis of language transcripts* (Research Version 8) [Computer software]. Madison, WI: University of Wisconsin—Madison.
- Miller, J.F., Long, S., McKinely, N., Thormann, S.S., Jones, M.A., & Nockerts, A. (2005). *Language sample analysis II: The Wisconsin guide*. Milwaukee, WI: Wisconsin Department of Public Instruction.
- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2006). *Mx: Statistical modeling* (7th ed.). Richmond, VA: Virginia Commonwealth University.
- Nippold, M. A. (1998). *Later language development: The school-age and adolescent years* (2nd ed.). Austin, Tx: Pro-Ed.
- Nippold, M. A. (2000). Language development during the adolescent years: Aspects of pragmatics, syntax, and semantics. *Topics in Language Disorders*, 20(2), 15-28. doi: 10.1097/00011363-200020020-00004
- Nippold, M. A., (2009). School-age children talk about chess: Does knowledge drive syntactic complexity?. *Journal of Speech, Language, and Hearing Research*, 52(4), 856-871. doi: 10.1044/1092-4388(2009/08-0094)
- Nippold, M. A., Hesketh, L. J., Duthie, J. K., & Mansfield, T. C. (2005). Conversational versus expository discourse: a study of syntactic development in children, adolescents, and adults. *Journal of Speech, Language, & Hearing Research*, 48(5), 1048-1064. doi: 10.1044/1092-4388(2005/073)

- Nippold, M. A., & Sun, L. (2008). Knowledge of morphologically complex words: A developmental study of older children and young adolescents. *Language, Speech, and Hearing Services in Schools, 39*(3), 365-373. doi: 10.1044/0161-1461(2008/034)
- Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, & Hearing Research, 57*(3), 876-886. doi: 10.1044/1092-4388(2013/13-0097)
- Oliver, B. R., & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics, 10*(01), 96-105. doi: 10.1375/twin.10.1.96
- Paul, R. (2001). Assessing advanced language. In R. Pauls (Ed.), *Language disorders from infancy through adolescence* (2nd ed, pp.501-546). St. Louis, MO: Mosby
- Peña, E., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*(2), 138-154. doi: 10.1044/1058-0360(2001/014)
- Peper, J. S., Schnack, H. G., Brouwer, R. M., Van Baal, G. C. M., Pjetri, E., Szekely, E., ... Hilleke, E. (2009). Heritability of regional and global brain structure at the onset of puberty: A magnetic resonance imaging study in 9-year-old twin pairs. *Human Brain Mapping, 30*(7), 2184-2196. doi: 10.1002/hbm.20660
- Petrill, S. A., Deater-Deckard, K., Thompson, L. A., DeThorne, L. S., & Schatschneider, C. (2006). Reading skills in early readers: Genetic and shared environmental effects. *Journal of Learning Disabilities, 39*(1), 48–55. doi : 10.1177/00222194060390010501

- Plomin, R., DeFries, J. C., & Loehlin, J. C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin*, 84(2), 309. doi: 10.1037/0033-2909.84.2.309
- Podsakoff, P.M., MacKenzie, S.B., Lee, J.-Y., & Podsakoff, N.P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. doi:10.1037/0021-9010.88.5.879
- Rice, D., & Barone Jr, S. (2000). Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. *Environmental Health Perspectives*, 108(Suppl 3), 511-533.
- Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, & Hearing Research*, 49(4), 793-808. doi: 10.1044/1092-4388(2006/056)
- Rogers, C.R., Nulty, K.L., Aparicio Betancourt, M. & DeThorne, L.S. (2015). Causal effects on child language development: A review of studies in communication sciences and disorders. *Journal of communication disorders*, 57, 3-15. doi: 10.1016/j.jcomdis.2015.06.004
- Rowe, D.C., Jacobson, K.C., & Van den Oord, E.J.C.G. (1999). Genetic and environmental influences on vocabulary IQ: Parental education level as moderator. *Child Development*, 70(5), 1151-1162. doi: 10.1111/1467-8624.00084
- Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype→ environment effects. *Child Development*, 54(2), 424-435. doi:

10.2307/1129703

- Scott, C.M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, & Hearing Research, 43*(2), 324-339. doi: 10.1044/jslhr.4302.324
- Semel, E., Wiig, E.H., & Secord, W.A. (2003). *Clinical evaluation of language fundamentals* (4th ed.). San Antonio, TX: PsychCorp.
- Speltz, M. L., DeKlyen, M., Calderon, R., Greenberg, M. T., & Fisher, P. A. (1999). Neuropsychological characteristics and test behaviors of boys with early onset conduct problems. *Journal of Abnormal Psychology, 108*(2), 315. doi:10.1037/0021-843X.108.2.315
- Spinath, F.M., Price, T.S., Dale, P.S., & Plomin, R. (2004). The genetic and environmental origins of language disability and ability. *Child Development, 75*(2), 445-454. doi: 10.1111/j.1467-8624.2004.00685.x
- Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences, 9*(2), 69-74. doi: 10.1016/j.tics.2004.12.005
- Stokes, S.F., Wong, A.M-Y., Fletcher, P., & Leonard, L.B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, & Hearing Research, 49*, 219-236. doi:10.1044/1092-4388(2006/019)
- Stromswold, K. (2001). The heritability of language: A review and metaanalysis of twin, adoption, and linkage studies. *Language, 77*(4), 647-723.
- Thompson, P. M., Giedd, J. N., Woods, R. P., MacDonald, D., Evans, A. C., & Toga, A. W.

- (2000). Growth patterns in the developing brain detected by using continuum mechanical tensor maps. *Nature*, 404(6774), 190-193. doi: 10.1038/35004593
- Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives potential indicators of language proficiency?. *Communication Disorders Quarterly*, 29(1), 43-53. doi: 10.1177/1525740108314866
- Ukrainetz, T. A., & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy*, 18(1), 59-78. doi: 10.1191/0265659002ct227oa
- Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, & Hearing Research*, 38(6), 1349-1355. doi: 10.1044/jshr.3806.1349
- Wiig, E.H., & Secord, W.A. (1989). *Test of language competence-expanded*. San Antonio, TX: Pearson.